A Graph Analysis of the Linked Data Cloud

Marko A. Rodriguez
Semantic Network Research Group
Knowledge Reef Systems Inc.
Santa Fe, New Mexico 87501
(Dated: March 2, 2009)

The Linked Data community is focused on integrating Resource Description Framework (RDF) data sets into a single unified representation known as the Web of Data. The Web of Data can be traversed by both man and machine and shows promise as the *de facto* standard for integrating data world wide much like the World Wide Web is the *de facto* standard for integrating documents. On February 27$^{\text{th}}$ of 2009, an updated Linked Data cloud visualization was made publicly available. This visualization represents the various RDF data sets currently in the Linked Data cloud and their interlinking relationships. For the purposes of this article, this visual representation was manually transformed into a directed graph and analyzed.

I. INTRODUCTION

The World Wide Web is a distributed document and media repository [1]. Hyper-Text Markup Language (HTML) documents reference other HTML documents and media (e.g. images, audio, etc.) by means of an `href` citation. The resulting document citation graph has been the object of scholastic research [2, 3] as well as a component utilized in web page ranking [4]. Similarly, the Semantic Web is a distributed resource identifier repository [5]. The Resource Description Framework (RDF) serves as one of the primary standards of the Semantic Web [6]. RDF provides the means by which Uniform Resource Identifiers (URI) [7] are interrelated to form a multi-relational or edge labeled graph. If U is the set of all URIs, L is the set of all literals, and B is the set of all blank (or anonymous) nodes, the the Semantic Web RDF graph is defined as the set of triples

$$G \subseteq (U \cup B) \times U \times (U \cup L \cup B).$$

Given that the URI is the foundational standard of both the World Wide Web and the Semantic Web, the Semantic Web serves as an extension to the World Wide Web in that it provides a semantically-rich graph overlay for URIs. Thus, the Semantic Web moves the Web beyond the simplistic `href` citation into a rich relational structure that can be utilized for numerous end user applications.

The Linked Data community is actively focused on integrating RDF data sets into a single connected data set [8]. The Linked Data model allows

> "[any man or machine] to start with one data source and then move through a potentially endless Web of data sources connected by RDF links. Just as the traditional document Web can be crawled by following hypertext links, the Web of Data can be crawled by following RDF links. Working on the crawled data, search engines can provide sophisticated query capabilities, similar to those provided by conventional relational databases. Because the query results themselves are structured data, not just links to HTML pages, they can be immediately processed, thus enabling a new class of applications based on the Web of Data." [9]

While the Linked Data community has focused on providing a distributed data structure, they have not focused on providing a distributed process infrastructure [10]. Unfortunately, if only a data structure is provided, then processing that data structure will lead to what has occurred with the World Wide Web: a commercial industry focused on downloading, indexing, and providing search capabilities to that data. For the problem space of keyword search, this model suffices. However, the RDF data model is much richer than the World Wide Web citation data model. If data must be downloaded to a remote machine for processing, then only so much of the Web of Data can be processed in a reasonable amount of time. This ultimately limits the sophistication of the algorithms that can be executed on the Web of Data. The RDF data model is rich enough to conveniently support the representation of relational objects [11] and their computational instructions [12]. Moreover, with respect to searching, the RDF data model requires a new degree of sophistication in graph analysis algorithms [13]. For one, the typical PageRank centrality calculation is nearly meaningless on an edge labeled graph [14]. To leave this algorithmic requirement to a small set of search engines will ultimately yield a limited set of algorithms and not a flourishing democracy of collaborative development. As a remedy to this situation, a distributed process infrastructure (analogous in many ways to the Grid [15]) may be a necessary requirement to ensure the accelerated, grass roots use of the Web of Data, where processes are migrated to the data, not data to the processes. In such a model, computational clock cycles are as open as the data upon which they operate.

With respect to the Web of Data as a distributed RDF data structure, this article presents a graph analysis of the March 2009 Linked Data cloud visualization that was published on February 27, 2009 by Chris Bizer.[24] The remainder of this article is organized as follows. §II articulates how the Linked Data cloud graph was constructed from the February 27$^{\text{th}}$ Linked Data cloud visualization.

§III provides a collection of standard graph statistics for the constructed Linked Data cloud graph. Finally §IV provides a more in-depth analysis of the structural properties of the graph.

II. CONSTRUCTING THE LINKED DATA CLOUD GRAPH

The current Linked Data cloud visualization was published by Chris Bizer on February 27, 2009. This visualization is provided in Figure 1. The Linked Data

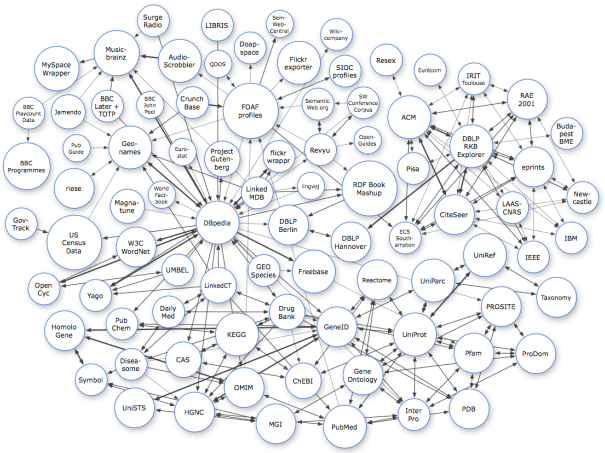

FIG. 1: The Linked Data cloud visualization as provided by the Linked Data community. This version is dated February 27, 2009. The author was not responsible for the creation of this visualization. This is only provided in order to better elucidate the means by which the Linked Data cloud graph was created.

cloud visualization represents various data sets as vertices (i.e. nodes) and their interlinking relationships as directed unlabeled edges (i.e. links). Moreover, it is assumed that vertex size denotes the number of triples in the data set and edge thickness denotes the extent to which one data set interlinks with another. Data set A links to data set B if data set B has a URI that is maintained (according to namespace) by data set A. In this way, by resolving a data set B URI within data set A, the man or machine is able to traverse to data set B from A.

A manual process was undertaken to turn the Linked Data cloud visualization into a Linked Data cloud graph denoted $G = (V, E)$, where V is the set of vertices (i.e. data sets), E is the set of unlabeled edges (i.e data set links), and $E \subseteq (V \times V)$. The link weights and the node sizes in the original visualization were ignored. A new visualization of the manually generated Linked Data cloud graph is represented in Figure 4. The properties of this visualization are discussed throughout the remainder of this article.

III. STANDARD GRAPH STATISTICS

Given the constructed Linked Data cloud graph visualized in Figure 4, it is possible to calculate various graph statistics. A collection of standard graph statistics are provided in Table I.

statistic	statistic value
number of vertices	86
number of edges	274
weakly connected	true
strongly connected	false
diameter	10
average path length	3.916

TABLE I: A collection of standard graph statistics for the Linked Data cloud graph represented in Figure 4.

A. Strongly Connected Components

The Linked Data graph is not strongly connected. This means that there does not exist a path from every data set to every other data set. Therefore, a walk along the graph can lead to an "island" of data sets that can not be returned from. The number of strongly connected components is 31 with 26 of those components only maintaining a single data set (that is, they are either the source of a path or the sink of a path). The size of the remaining strongly connected components is 37, 15, 4, 2, and 2. The largest component (with size of 37) is the "DBpedia component". The second largest (with size of 15) is the "DBLP RKB Explorer component".

Given the large diameter and average path length, the Linked Data cloud graph can be seen as a two weakly connected components: the larger DBpedia component and the smaller DBLP RKB Explorer component. However, as will be seen later, other communities in the larger DBpedia component exist such as biological and medical communities.

B. Degree Distributions

The in- and out-degree distributions of the graph are plotted in Figure 2 and Figure 3 on a log-log plot, respectively. These plots show the number (frequency) of data sets that have a particular in- or out-degree. The top 11 in- and out-degree data sets are presented in Table II and Table III, respectively. It is interesting to note that the two leaders (DBpedia and DBLP RKB Explorer) are also the leaders of the two largest strongly connected components identified previously.

While the number of data points is small, a power-law fit is provided according to a distribution that is defined as $p(x) \sim x^{-\alpha}$, where $p(x)$ is the probability of seeing a data set with a degree of x. A power-law fit to the total degree distribution (i.e. ignoring edge directionality)

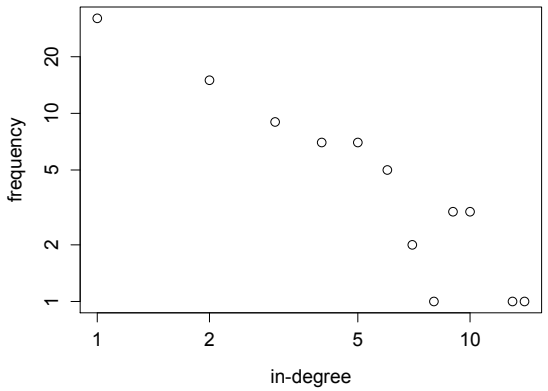

FIG. 2: The in-degree distribution of the Linked Data cloud graph on a log-log plot.

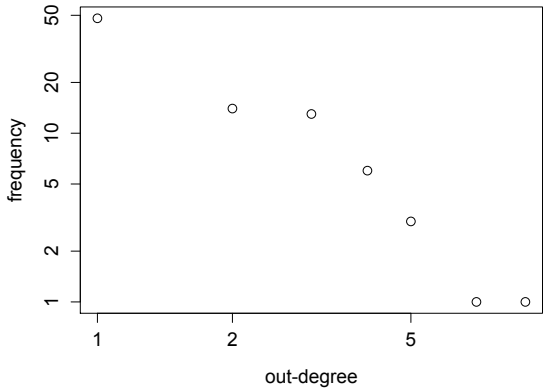

FIG. 3: The out-degree distribution of the Linked Data cloud graph on a log-log plot.

data set	in-degree
DBpedia	14
DBLP RKB Explorer	13
ACM	10
GeneID	10
GeoNames	10
CiteSeer	9
ePrints	9
UniProt	9
ECS Southampton	8
FOAF Profiles	7
RAE 2001	7

TABLE II: The top 11 Linked Data data sets with the highest in-degree.

yields an exponent of $\alpha = 1.496$. In other words, the larger the degree, the fewer number of data sets.

C. Degree Correlations

The correlation between the in- and out-degrees of the vertices yields a Spearman $\rho = 0.6753$ with a significant $p < 9.85^{-13}$. Similarly, the Kendall $\tau = 0.5640$ with a significant $p < 7.27^{-12}$. In other words, data sets that frequently link to other data sets tend to get linked to frequently.

data set	out-degree
DBpedia	17
DBLP RKB Explorer	14
ACM	10
CiteSeer	9
EPrints	9
GeneID	8
UniProt	8
DrugBank	7
ECS Southampton	7
FOAF Profiles	7
RAE 2001	7

TABLE III: The top 11 Linked Data data sets with the highest out-degree.

If a graph is degree assortative then vertices with high degree are connected to other vertices with high degree. Likewise, vertices with low degree connect to vertices with low degree. Assortativity is calculated by creating two vectors of length $|E|$. One vector maintains the degree of the vertices at the head of each edge and the other vector maintains the degree of the vertices at the tail of each edge. These two vectors are then correlated. The popular assortative mixing value [16] is calculated with a Pearson correlation over the two vectors as

$$r = \frac{|E| \sum_i j_i k_i - \sum_i j_i \sum_i k_i}{\sqrt{\left[|E| \sum_i j_i^2 - (\sum_i j_i)^2\right] \left[|E| \sum_i k_i^2 - (\sum_i k_i)^2\right]}},$$

where j_i is the degree of the vertex on the tail of edge i, and k_i is the degree of the vertex on the head of edge i. The correlation coefficient r is in $[-1, 1]$, where -1 represents a fully disassortative graph, 0 represents an uncorrelated graph, and 1 represents a fully assortative graph. Given that the degree distribution is non-parametric, a non-parametric assortativity correlation is also provided using both Spearman ρ and Kendall τ. All of these assortativity correlations are presented in Table IV, where the only significant values are from the standard Pearson correlation and all the in-degree correlations. These results demonstrate that Linked Data data sets tend to connect to data sets with differing degrees. That is, for instance, high degree data sets connect to low degree data sets. This is made apparent when looking at DBpedia

method	in-degree	out-degree	total-degree
pearson	-0.1911 (0.0015)	-0.1728 (0.0042)	-0.1868 (0.0019)
spearman	-0.1319 (0.0292)	-0.0311(0.6089)	-0.0629 (0.2998)
kendall	-0.0933 (0.0346)	-0.0193 (0.6626)	-0.0364 (0.3982)

TABLE IV: Various degree assortativity correlations for the Linked Data cloud graph. The first number is the correlation and the second number in parentheses is the p-value. A significant p-value is less than 0.05.

which has a total-degree of 32. DBpedia's neighbors in the graph have the following total-degrees: 1, 1, 2, 3, 3, 3, 3, 3, 4, 4, 4, 6, 8, 9, 11, 12, 12, and 18. However, in general, the degree assortativity correlation is weak and for the non-parametric correlations, mostly insignificant.

IV. STRUCTURAL ANALYSIS

This section presents an analysis of the community structures that exist within the Linked Data cloud graph. A community is loosely defined as a set of vertices that have a high number of intra-connections and low number of inter-connections. In other words, vertices in the same community tend to link to vertices in the same community as opposed to vertices in other communities. In order to compare the algorithmically determined structural communities to the metadata properties of the vertices that compose those communities, two metadata properties were gathered:

1. a string label denoting the type of content maintained in the data set
2. an integer value denoting the number of triples contained in the data set.

The content labels were determined manually. The set of labels used were: biology, business, computer science, general, government, images, library, location, media, medicine, movie, music, reference, and social. Note that many data sets could have been labeled with more than one label. However, only one label was chosen. Moreover, these labels were determined by reviewing the websites of the data sets and not by looking at the structure of the graph.

The data set triple counts were taken from the "Linking Open Data on the Semantic Web" web page.[25] Of the 86 data sets in the Linked Data cloud, only 31 of those data sets have published triple counts.

A. Labeled Structural Communities

The graph analysis method for comparing nominal vertex metadata with structural communities as originally presented in [17] was used to compare the content labels of the data sets to their structural communities. The purpose of this analysis is to determine the semantics of the structural communities. The hypothesis is that structural communities denote shared content. That is, data sets in the same structural community maintain the same type of content data (e.g. biology, medicine, computer science, etc.).

A contingency table was created that denotes the number of vertices that have a particular content label and are in a particular structural community. An example contingency table that has community values that were determined using the leading eigenvector community detection algorithm [18] is presented in Table V. The contingency table is subjected to a χ^2 analysis in order to determine if the manually generated content labels are statistically related to the algorithmically determined structural communities. Four community detection algorithms (and thus, four individual contingency tables) were used for this analysis and the χ^2 p-values are presented in Table VI.

content/community	0	1	2	3	4	5	6	7	8	9
biology	2	0	4	1	0	0	0	0	3	10
business	1	0	0	0	0	1	0	0	0	0
computer science	1	12	0	0	0	0	0	2	0	0
general	4	3	0	0	0	0	0	1	0	0
government	3	0	0	0	0	2	0	0	0	0
images	1	0	0	0	0	0	0	1	0	0
library	2	0	0	0	0	1	0	1	0	0
location	0	0	0	0	0	1	0	1	0	0
media	0	0	0	0	0	0	1	0	0	0
medicine	0	1	0	4	0	0	0	0	1	1
movie	1	0	0	0	0	0	0	0	0	0
music	5	0	0	0	0	1	1	1	0	0
reference	2	0	1	1	0	0	0	0	0	0
social	0	0	0	0	1	0	0	5	0	1

TABLE V: An example contingency table that denotes how many data sets have a particular content label and structural community. For this example, the structural communities were determined using the leading eigenvector community detection algorithm.

community algorithm	χ^2 p-value
Leading Eigenvector	6.6^{-12}
WalkTrap	2.2^{-16}
Edge Betweenness	0.0323
Spinglass	2.4^{-16}

TABLE VI: The p-values for four χ^2 tests using four structural community detection algorithms: leading eigenvector [18], walktrap [19], edge betweenness [20], and spinglass [21].

The analysis demonstrates that data sets that maintain similar content tend to exist in the same structural areas of the graph. This is made salient by a qualitative analysis of various subsets of the graph (see Figure 4 where the vertex colors denote their structural community). Moreover, this makes sense intuitively. Data sets that share the same content labels are more than likely to reference to the same resources. For example, it is true that medical data sets tend to be connected to other medical data sets and not to music data sets. Table VII provides a review of 15 randomly chosen Linked Data data sets, their structural community values according to the leading eigenvector community detection algorithm, and their manually determined content labels.

B. Data Set Triple Counts

Of the 86 data sets in the Linked Data cloud, only 31 of those data sets have triple counts that were published

data set	community	content label
SurgeRadio	0	music
MusicBrainz	0	music
DBpedia	0	general
Riese	5	government
LinkedCT	3	medicine
World Fact Book	5	government
OpenCyc	0	general
Yago	0	general
DrugBank	3	medicine
DailyMed	3	medicine
UniParc	2	biology
Reactome	9	biology
ACM	1	computer science
CiteSeer	1	computer science
IEEE	1	computer science

TABLE VII: A sample of 15 Linked Data data sets, their leading eigenvector structual community value, and their manually determined content label.

on the "Linking Open Data on the Semantic Web" web page. Given the statistically significant, positive correlation between the in-degree and out-degree of the vertices, it is hypothesized that those data sets that are more central in the graph will have a larger triple count. The centrality of all 86 vertices was determined using the PageRank centrality algorithm with a $\delta = 0.85$ [22]. For those 31 data sets that have triple counts, their triple count value was correlated with their PageRank centrality value. The Spearman $\rho = 0.6274$ with a significant $p < 0.00016$. Similarly, the Kendall $\tau = 0.4566$ with a significant $p < 0.00039$. Thus, those data sets that have the most RDF triples tend to be centrally located in the Linked Data cloud.

Finally, an assortative mixing calculation over data set triple counts was performed. Given that only 31 data sets have triple count values, a 31 vertex subgraph was created. This 31 vertex graph has 56 edges. These 56 edges were used to determine the assortative triple count correlation. Thus, two vectors of length 56 were created where one vector maintained the triple count of the data sets on the head of each edge and the other vector maintained the triple count of the data sets on the tail of each edge. Table VIII provides three assortativity correlations. Note that the triple count data distribution is non-parametric. From these results, only the non-parametric Kendall correlation is statistically significant with a correlation that demonstrates that the data sets are loosely disassortative according. This means that small data sets tend to connect to large data sets and large data sets tend to connect to small data sets. Again, this correlation is relatively weak.

C. Data Set Centrality

The PageRank centrality (with $\delta = 0.85$) of each of the 86 data sets in the Linked Data cloud graph was calculated. Table IX provides the top 15 central data sets. From this analysis, and assuming that centrality denotes "importance", it appears that content in computer science and biology are of major import to the current instantiation of the Linked Data cloud.

method	size assortativity
pearson	0.0682 (0.3230)
spearman	-0.2546 (0.0559)
kendall	-0.2064 (0.0302)

TABLE VIII: Data set triple count assortativity correlations for the Linked Data cloud graph. Given that only 31 data sets have published triple counts, these assortativity values are determined according to this 31 data set subgraph. The first number is the correlation and the second number in parentheses is the p-value. A significant p-value is less than 0.05.

data set	page rank	content label
DBLP Berlin	0.0484	computer science
DBLP Hannover	0.0464	computer science
DBpedia	0.0384	general
KEGG	0.0370	biology
UniProt	0.0357	biology
GeneID	0.0346	biology
DBLP RKB Explorer	0.0341	computer science
GeoNames	0.0294	location
ACM	0.0257	computer science
Pfam	0.0254	biology
Prosite	0.0233	biology
ePrints	0.0218	computer science
CiteSeer	0.0218	computer science
PDB	0.0209	biology

TABLE IX: The top 15 PageRank central data sets in the Linked Data cloud graph.

V. CONCLUSION

The Linked Data initiative is focused on unifying RDF data sets into a single global data set that can be utilized by both man and machine. This initiative is providing a fundamental shift in the way in which data is maintained, exposed, and interrelated. This shift is both technologically and culturally different from the relational database paradigm. For one, the address space of the Web of Data is the URI address space, which is inherently distributed and infinite. Second, the graph data structure is becoming a more accepted, flexible representational medium and as such, may soon displace the linked table data structure of the relational database model. Finally, with respects to culture, the Web of Data maintains publicly available interrelated data. In the relational database world, rarely are database ports made publicly available for harvesting and rarely are relational schemas published for reuse. The Semantic Web, the Linked Data community, and the Web of Data are truly emerging as a radical

rethinking of the way in which data is managed and distributed in the modern world.

[1] T. Berners-Lee, R. Cailliau, A. Luotonen, H. Nielsen, and A. Secret, "The World-Wide Web," *Communications of the ACM*, vol. 37, pp. 76–82, 1994.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," in *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands, May 2000.

[3] B. A. Huberman and L. A. Adamic, "Growth dynamics of the world-wide web," *Nature*, vol. 399, 1999.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[5] T. Berners-Lee and J. A. Hendler, "Publishing on the Semantic Web," *Nature*, vol. 410, no. 6832, pp. 1023–1024, April 2001. [Online]. Available: http://dx.doi.org/10.1038/35074206

[6] E. Miller, "An introduction to the Resource Description Framework," *D-Lib Magazine*, May 1998. [Online]. Available: http://dx.doi.org/hdl:cnri.dlib/may98-miller

[7] T. Berners-Lee, , R. Fielding, D. Software, L. Masinter, and A. Systems, "Uniform Resource Identifier (URI): Generic Syntax," January 2005.

[8] T. Berners-Lee, "Linked data," World Wide Web Consortium, Tech. Rep., 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[9] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the web," in *Proceedings of the International World Wide Web Conference*, ser. Linked Data Workshop, Beijing, China, April 2008.

[10] M. A. Rodriguez, "A distributed process infrastructure for a distributed data structure," *Semantic Web and Information Systems Bulletin*, 2008. [Online]. Available: http://arxiv.org/abs/0807.3908

[11] E. Oren, B. Heitmann, and S. Decker, "ActiveRDF: Embedding semantic web data into object-oriented languages," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 3, pp. 191–202, 2008.

[12] M. A. Rodriguez, *Emergent Web Intelligence*. Berlin, DE: Springer-Verlag, 2008, ch. General-Purpose Computing on a Semantic Network Substrate. [Online]. Available: http://arxiv.org/abs/0704.3395

[13] B. Aleman-Meza, C. Halaschek-Wiener, I. B. Arpinar, C. Ramakrishnan, and A. P. Sheth, "Ranking complex relationships on the semantic web," *IEEE Internet Computing*, vol. 9, no. 3, pp. 37–44, 2005.

[14] M. A. Rodriguez, "Grammar-based random walkers in semantic networks," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 727–739, 2008. [Online]. Available: http://arxiv.org/abs/0803.4355

[15] I. Foster and C. Kesselman, *The Grid*. Morgan Kaufmann, 2004.

[16] M. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, 2002.

[17] M. A. Rodriguez and A. Pepe, "On the relationship between the structural and socioacademic communities of an interdisciplinary coauthorship network," *Journal of Informetrics*, vol. 2, no. 3, pp. 195–201, July 2008. [Online]. Available: http://arxiv.org/abs/0801.2345

[18] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, May 2006. [Online]. Available: http://arxiv.org/abs/physics/0605087

[19] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, 2006.

[20] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, p. 7821, 2002.

[21] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 016110, 2006. [Online]. Available: http://arxiv.org/abs/cond-mat/0603718

[22] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.

[23] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[24] The March 2009 Linked Data cloud visualization is available at: `http://tinyurl.com/b4vfbq`.

[25] Linking Open Data on the Semantic Web is available at: `http://tinyurl.com/5fcmzm`.

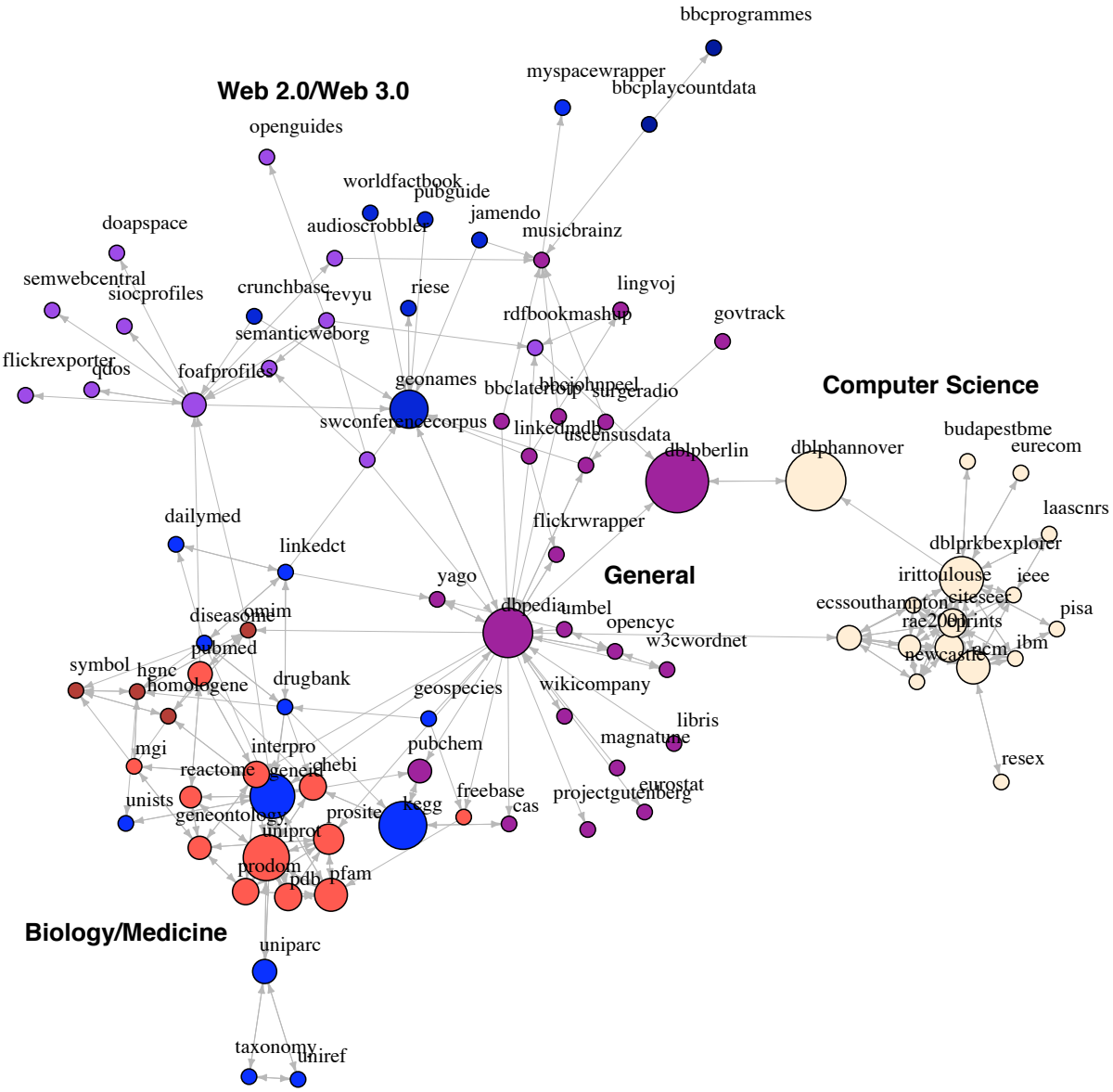

FIG. 4: A graph representation of the March 2009 Linked Data Cloud. Each vertex denotes a Linked Data data set. Each edge denotes whether one data set makes reference to another. The size of the vertices are determined by their PageRank centrality according to a $\delta = 0.85$ [22]. The vertex colors denote the structural communities as identified by the leading eigenvector community detection algorithm [18]. Finally, the Fruchterman-Reingold layout algorithm was used to visually render this representation [23].